下一代RAG引擎
—— 技术挑战与实现

演讲人：张颖峰

InfiniFlow/ 创始人

AiCon
全球人工智能开发与应用大会

InfoQ 极客传媒

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

4月 — QCon 北京 · 全球软件开发大会 · 会议时间：已结束

5月 — AiCon 北京 · 全球人工智能开发与应用大会 · 会议时间：已结束

6月 — ArchSummit 深圳 · 全球架构师峰会 · 会议时间：已结束

8月 — FCon 上海 · 全球金融科技大会 · 会议时间：8月16–17日

8月 — AiCon 上海 · 全球人工智能开发与应用大会 · 会议时间：8月18–19日

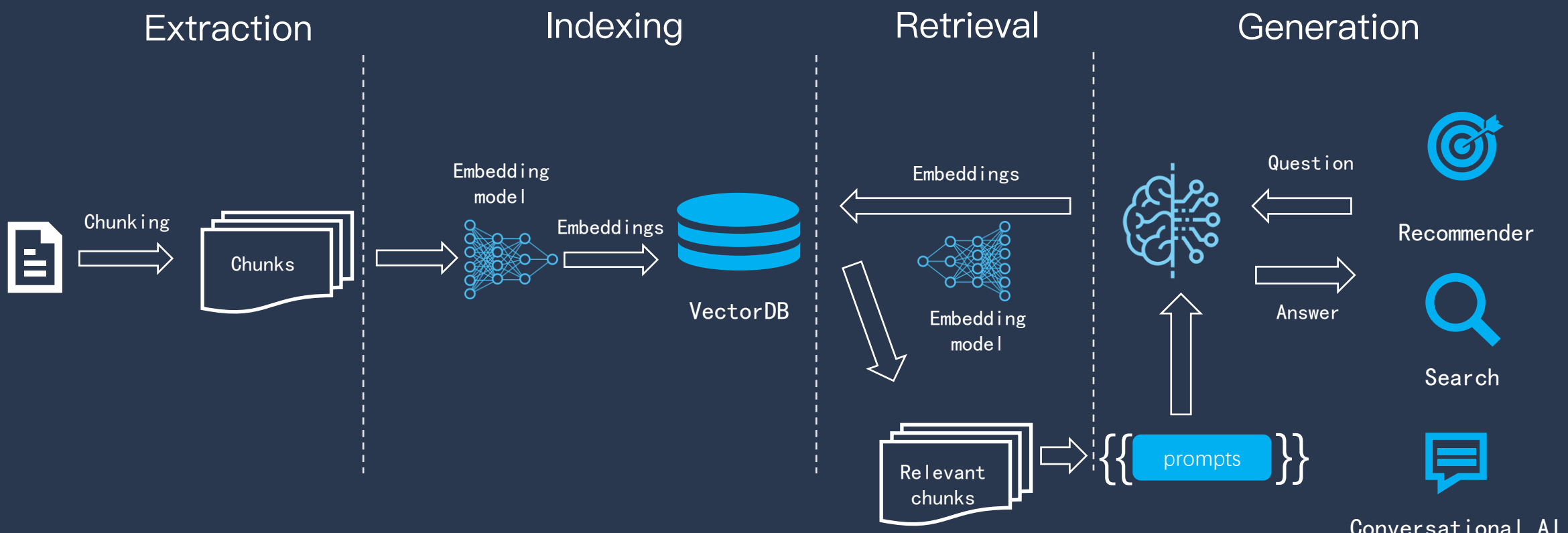
10月 — QCon 上海 · 全球软件开发大会 · 会议时间：10月18–19日

12月 — ArchSummit 北京 · 全球架构师峰会 · 会议时间：12月20–21日

目 录

01 下一代RAG引擎

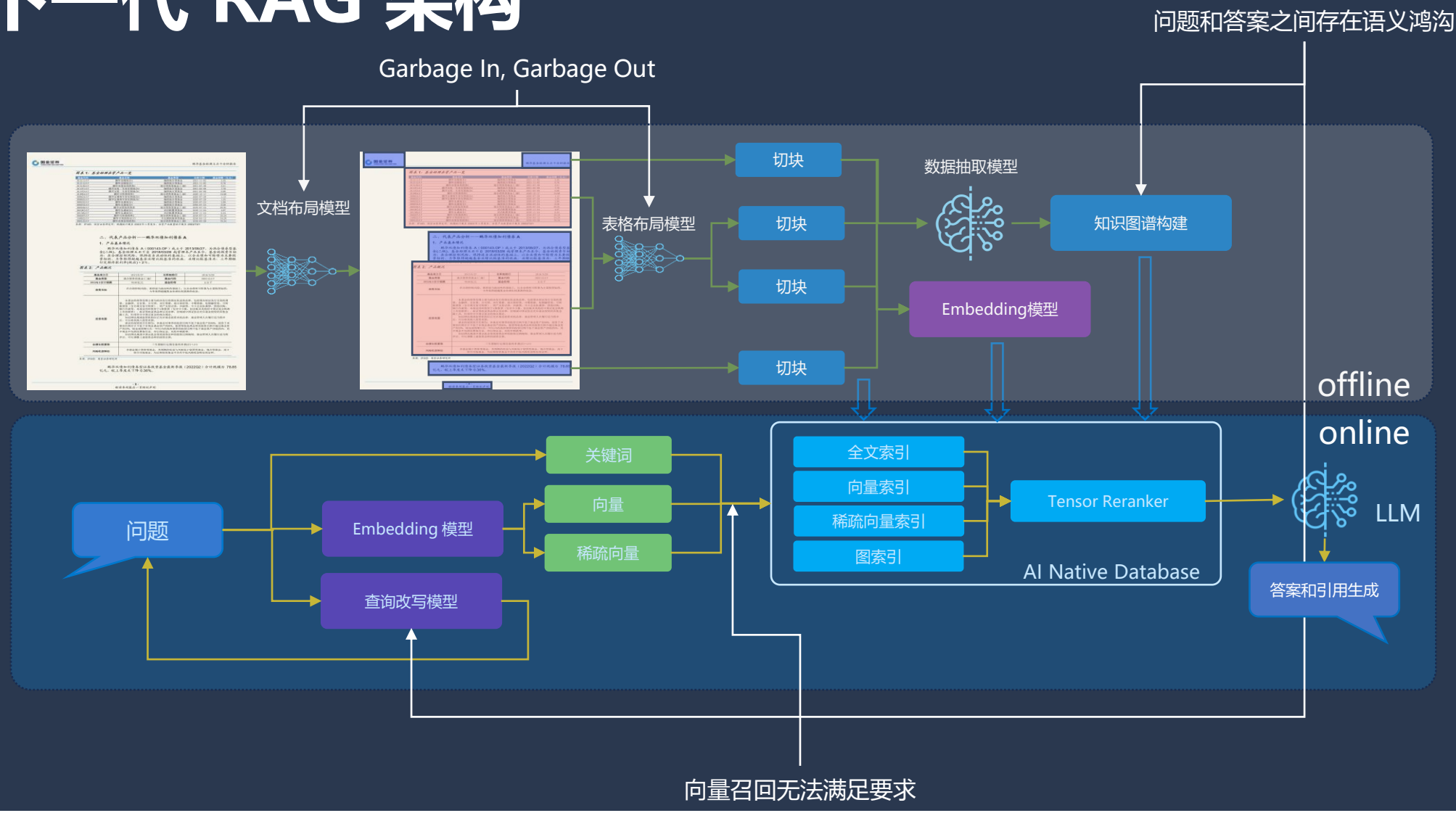
RAG 架构模式


Extraction
Indexing
Retrieval
Generation
Chunking
Chunks
Embedding model
Embeddings
VectorDB
Embeddings
Embedding model
Question
Answer
Recommender
Search
Conversational AI
Relevant chunks
prompts

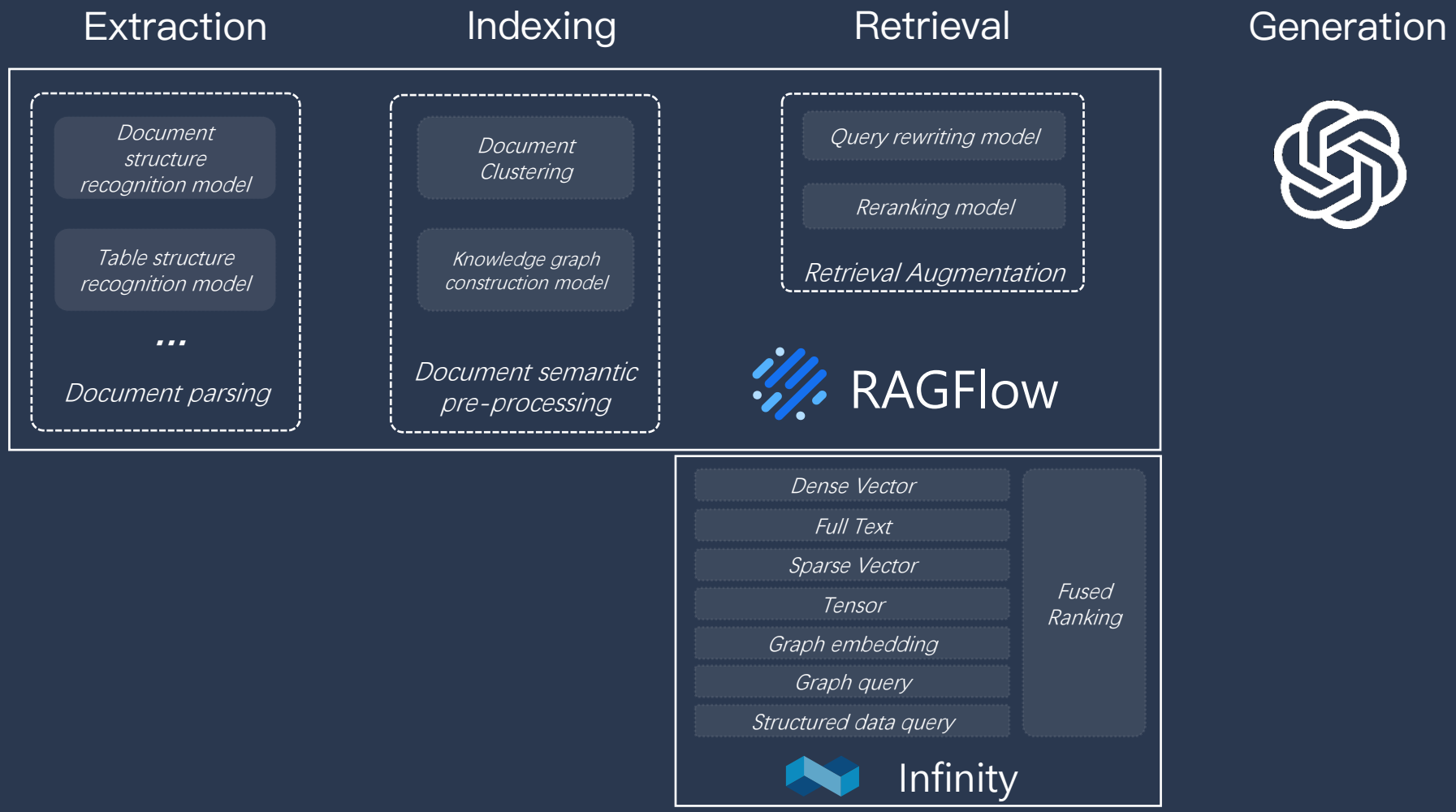
当前 RAG 面临的挑战

- 挑战一：向量的召回无法满足要求
- 挑战二：文档结构复杂，数据太乱， Garbage In, Garbage Out
- 挑战三：问题和答案所在文档关联不大，很难通过问题找到正确文档

下一代 RAG 架构


问题和答案之间存在语义鸿沟
Garbage In, Garbage Out
文档布局模型
表格布局模型
切块
切块
切块
切块
数据抽取模型
知识图谱构建
Embedding模型
offline
online
问题
Embedding 模型
查询改写模型
关键词
向量
稀疏向量
全文索引
向量索引
稀疏向量索引
图索引
Tensor Reranker
AI Native Database
LLM
答案和引用生成
向量召回无法满足要求

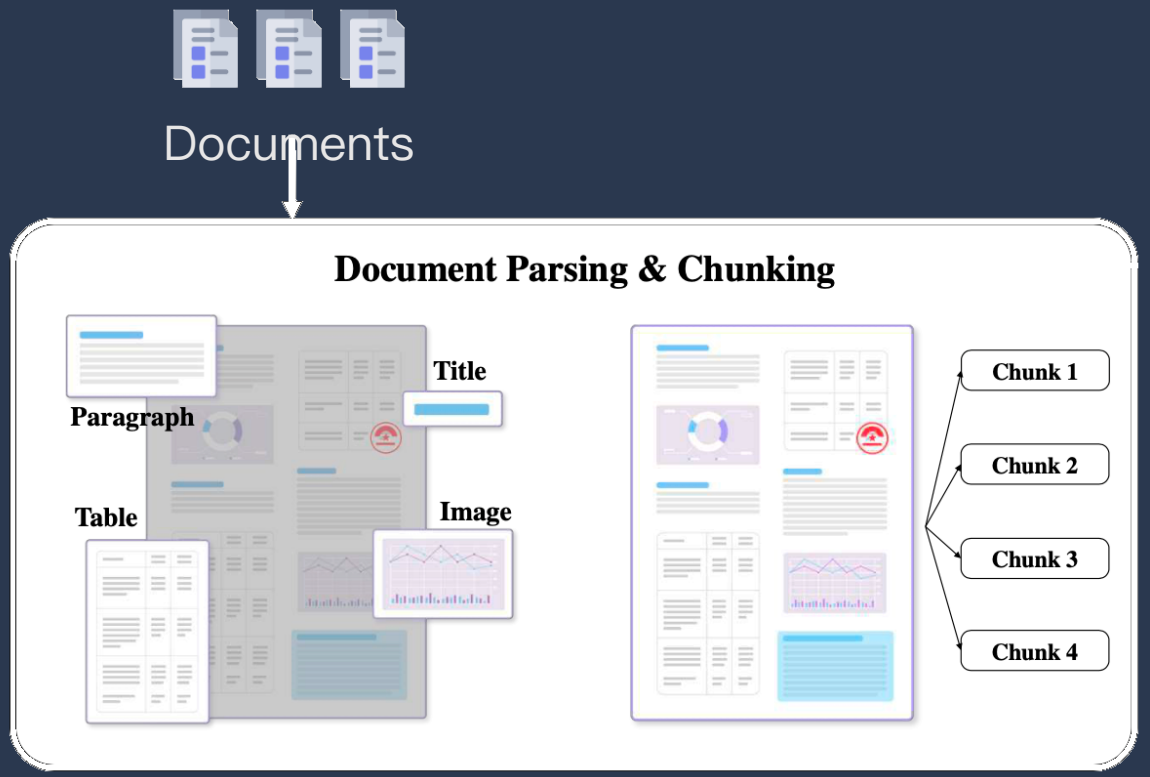
Infinity + RAGFlow = Infiniflow

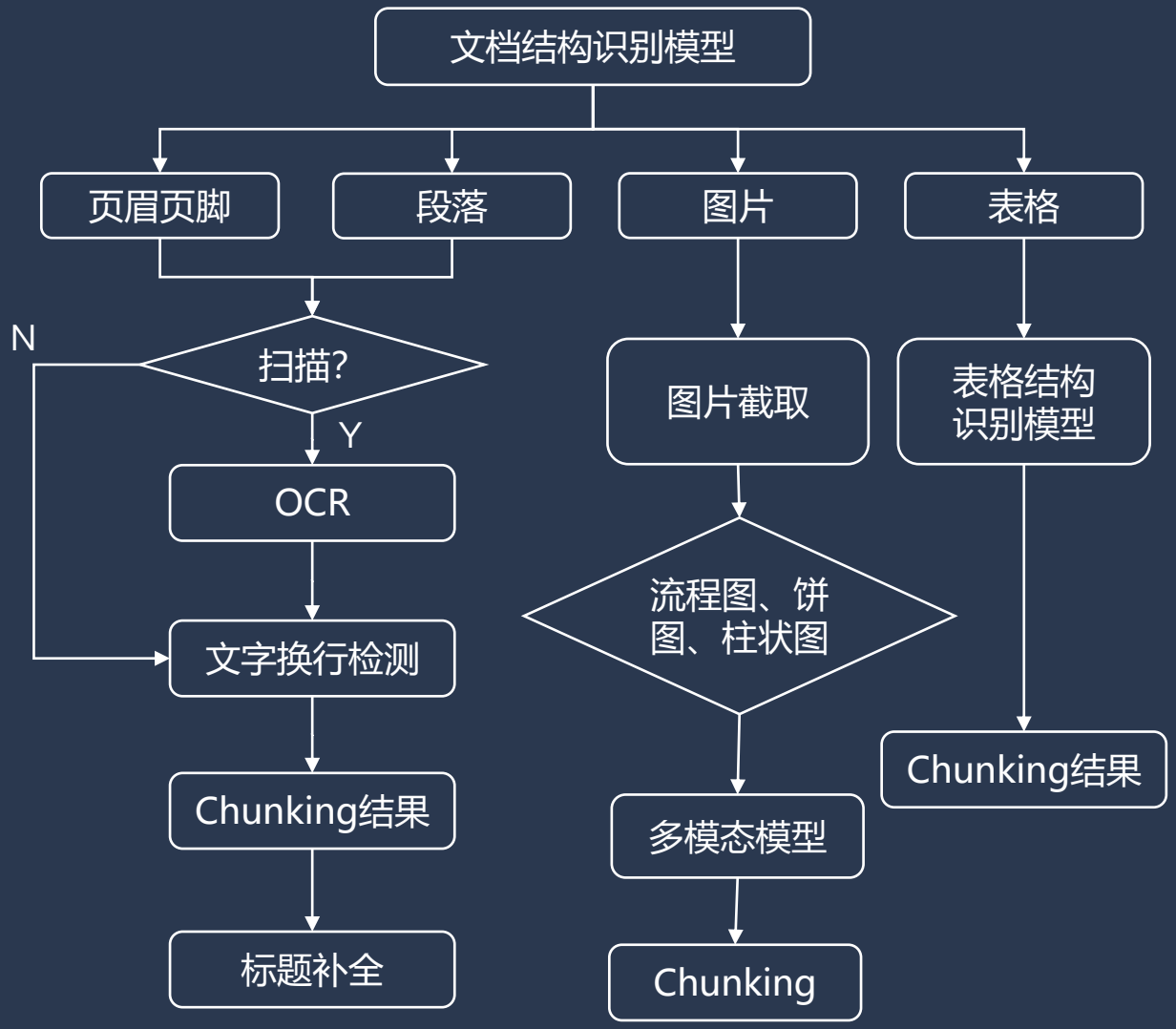


02

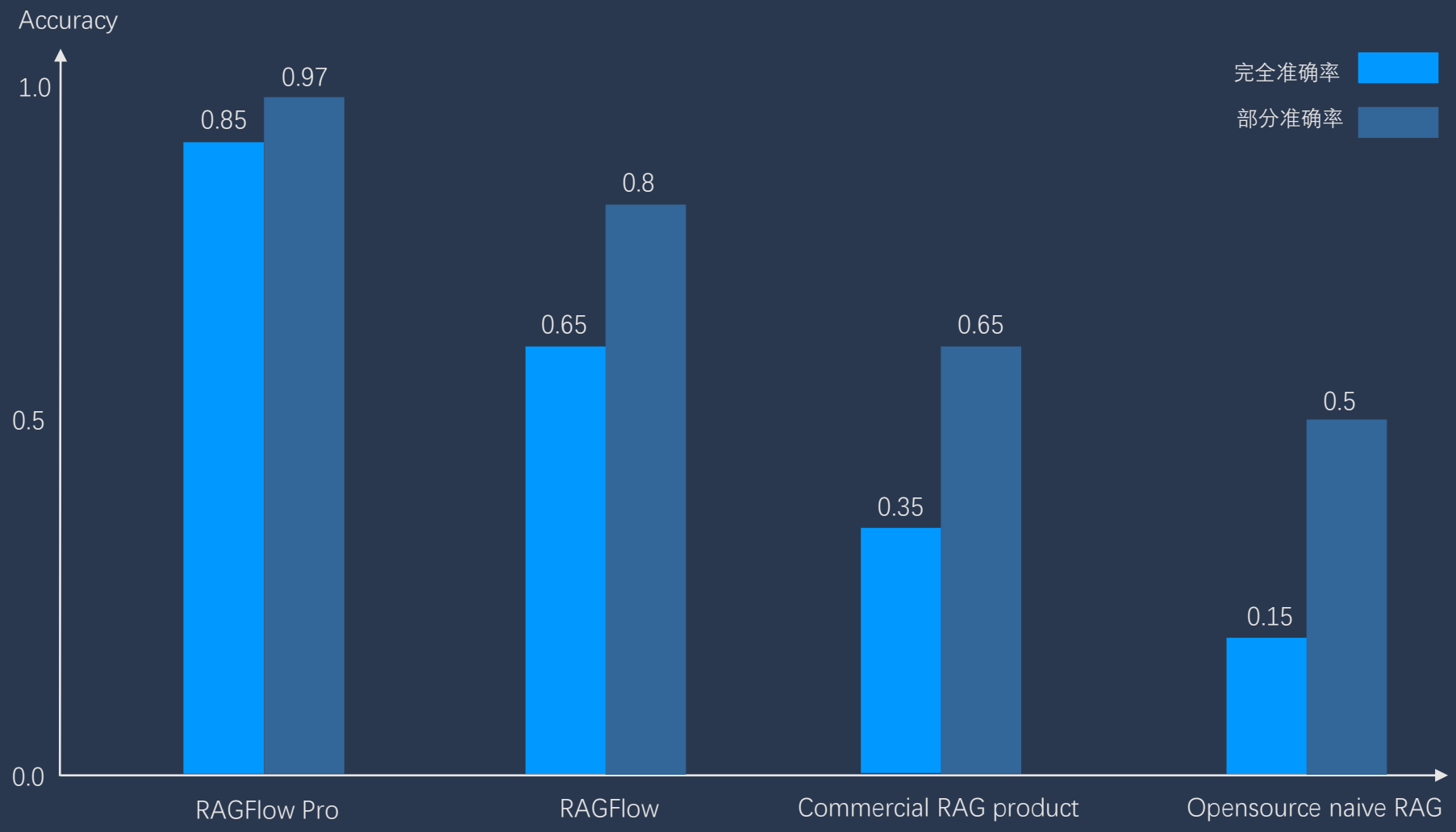
数据抽取模型

概要


Documents
Document Parsing & Chunking
Paragraph
Title
Table
Image
Chunk 1
Chunk 2
Chunk 3
Chunk 4


文档结构识别模型
页眉页脚
段落
图片
表格
N
扫描?
Y
OCR
文字换行检测
Chunking结果
标题补全
图片截取
流程图、饼图、柱状图
多模态模型
Chunking
表格结构识别模型
Chunking结果

调整抽取模型的 RAGFlow 对比



表格识别模型

板块	申万一级行业	2020Q 2	2020Q 4	2021Q 2	2021Q 4
制造	电力设备	14.37%	8.96%	15.63%	7.80%
	机械设备	3.03%	4.13%	4.54%	1.03%
	汽车	1.68%	2.45%	4.18%	1.33%
	公用事业	0.00%	0.00%	0.00%	5.32%
	环保	1.98%	0.96%	0.00%	2.67%
	国防军工	0.00%	7.42%	2.16%	6.08%
	综合	0.00%	0.00%	0.00%	0.00%
消费	纺织服饰	0.00%	0.00%	0.00%	0.00%
	家用电器	0.00%	1.06%	0.00%	0.90%
	农林牧渔	3.56%	0.00%	4.68%	4.10%
	轻工制造	2.95%	2.06%	0.86%	0.00%
	商贸零售	2.69%	1.25%	3.64%	0.15%
	食品饮料	7.90%	4.22%	0.83%	10.08%
	社会服务	1.53%	0.30%	1.03%	0.00%
	美容护理	0.00%	0.00%	2.20%	0.35%
医药	医药生物	15.64%	12.52%	9.55%	5.98%
科技	传媒	3.44%	0.00%	0.00%	2.50%
	电子	11.32%	9.25%	17.68%	11.48%
	计算机	11.21%	8.33%	2.62%	1.80%
	通信	2.70%	2.08%	0.00%	0.00%
金融	房地产	0.00%	0.00%	0.00%	0.70%
	非银金融	2.88%	8.31%	2.23%	9.34%
	银行	0.00%	4.18%	9.73%	0.36%
周期	煤炭	0.00%	0.00%	2.57%	0.00%
	石油石化	0.00%	1.15%	3.00%	0.00%
	钢铁	0.00%	3.45%	3.13%	3.46%
	基础化工	5.56%	3.37%	2.64%	9.78%
	建筑材料	3.09%	5.49%	4.20%	0.00%
	有色金属	2.50%	5.10%	2.09%	4.37%
	建筑装饰	0.00%	1.30%	0.00%	3.63%
	交通运输	1.97%	2.68%	0.84%	6.79%

- 单元格边界判定
- 表头信息判定
- 单元格合并判定
- 表格跨页判定

表格识别模型

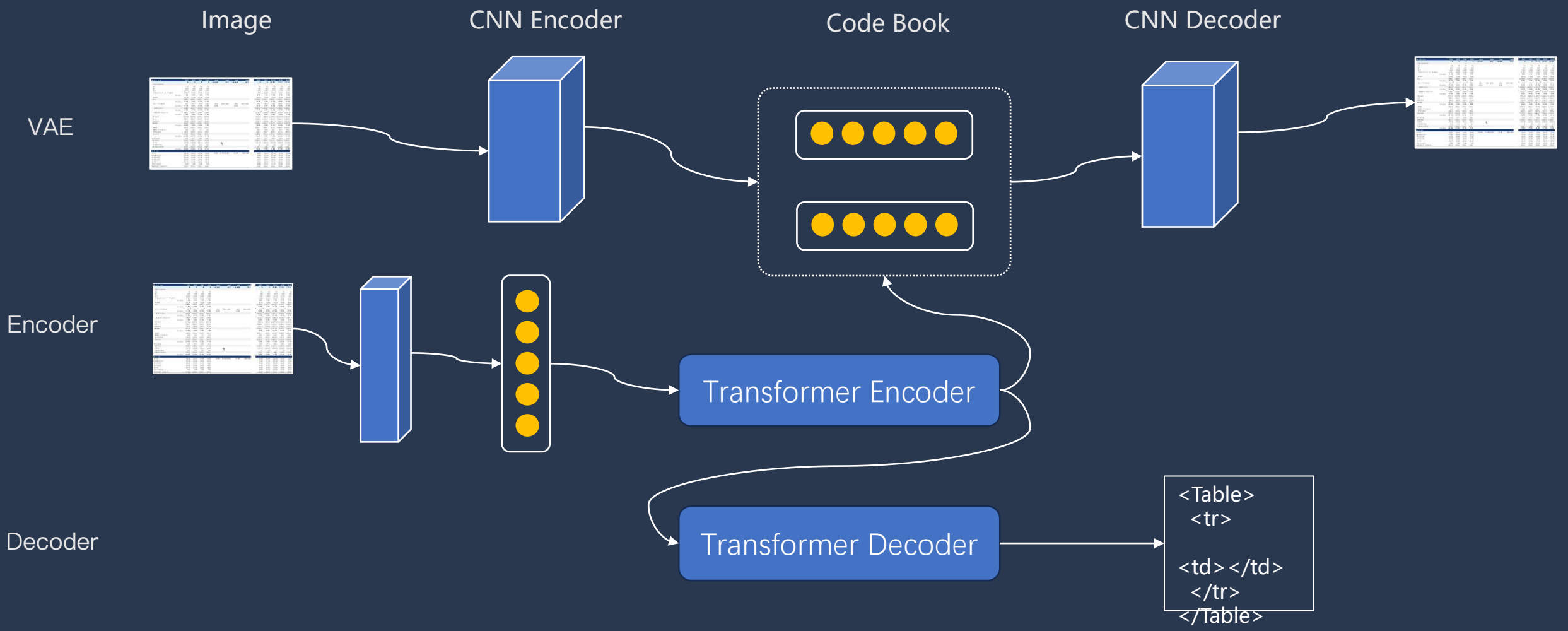
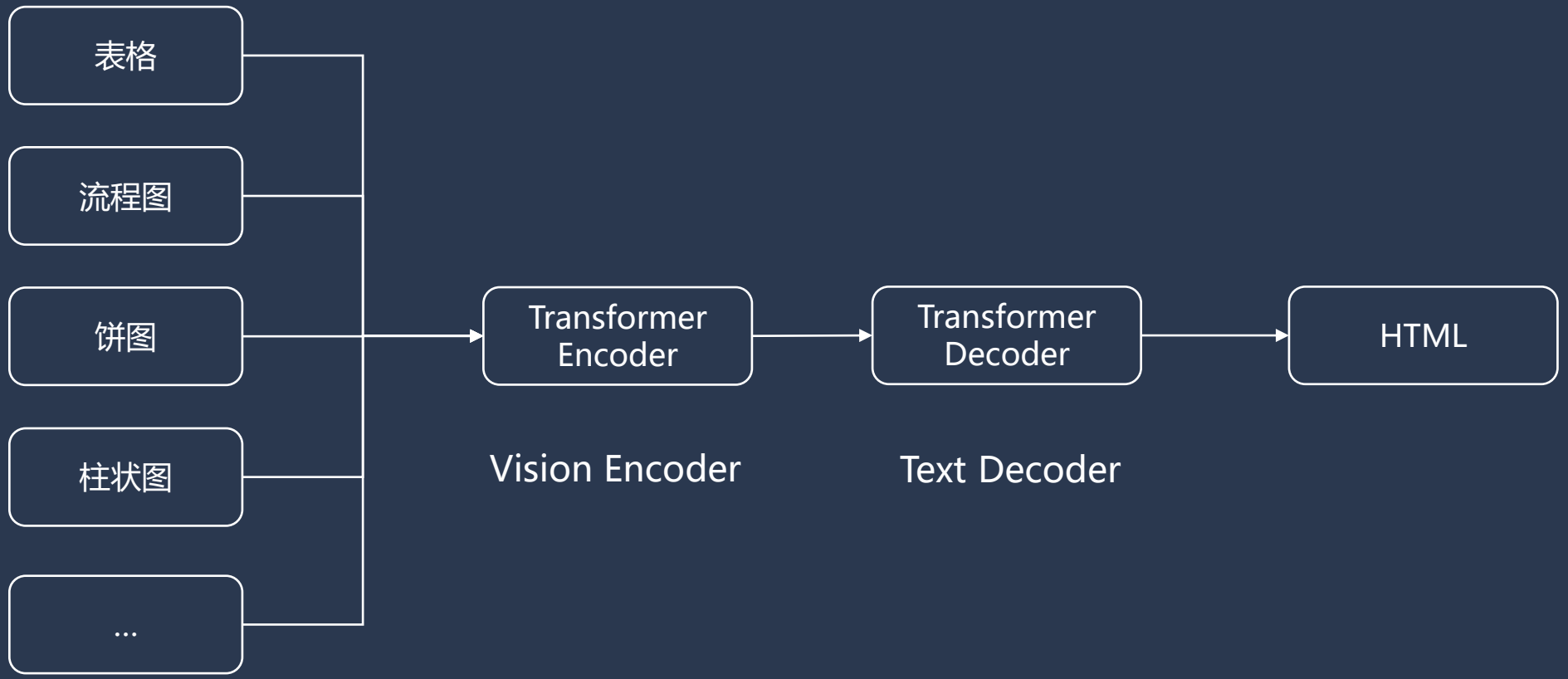
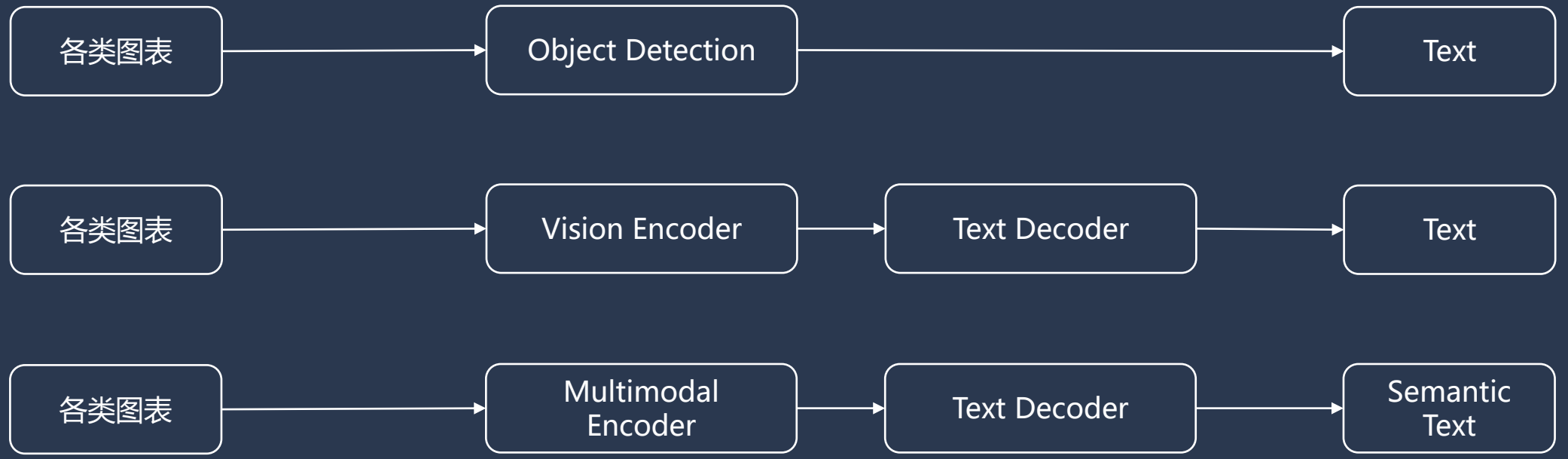


Image
CNN Encoder
Code Book
CNN Decoder
VAE
Encoder
Transformer Encoder
Transformer Decoder
Decoder
<Table>
<tr>
<td></td>
</tr>
</Table>

文档“大”模型


表格
流程图
饼图
柱状图
...
Transformer Encoder
Transformer Decoder
HTML
Vision Encoder
Text Decoder

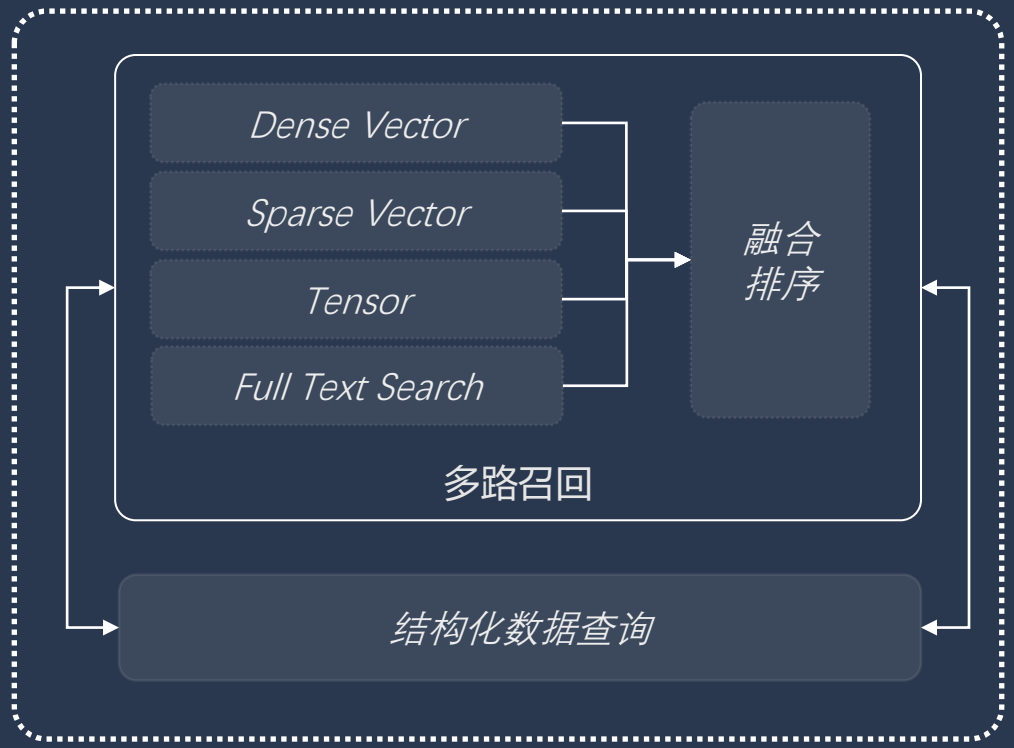
“雕花”还是多模态LLM?

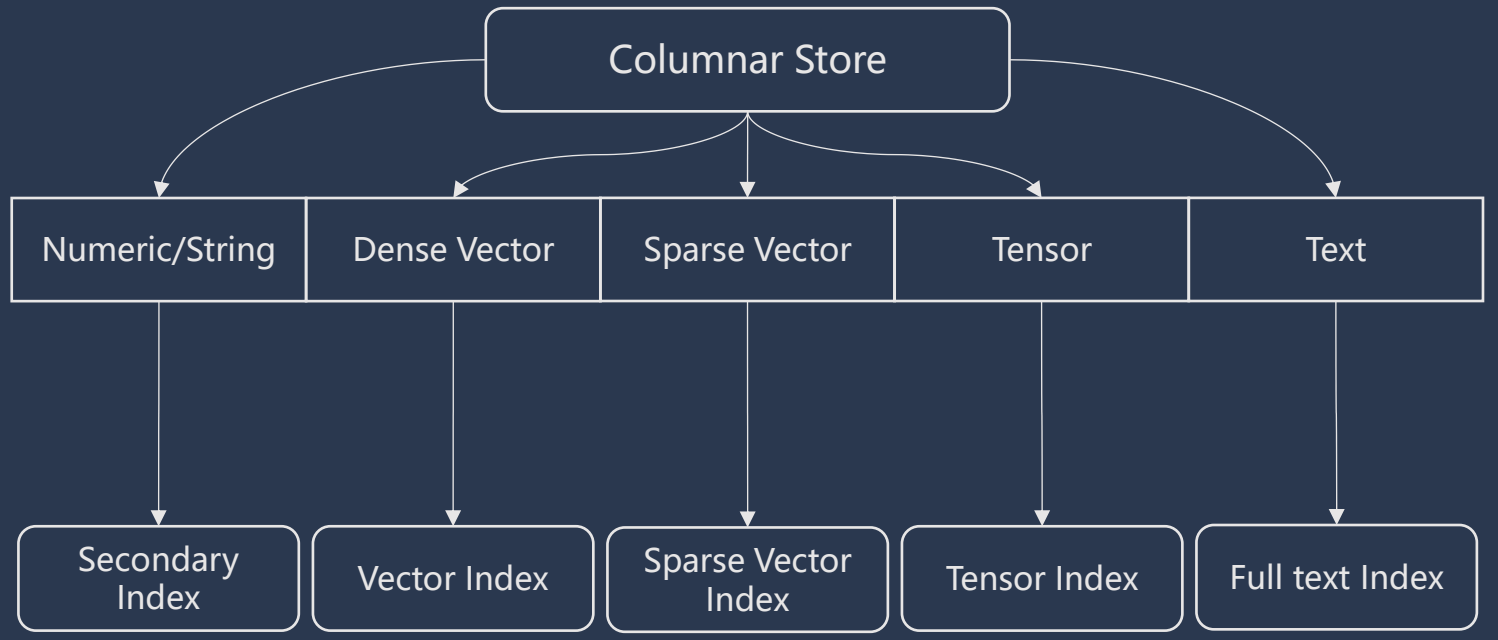


CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs

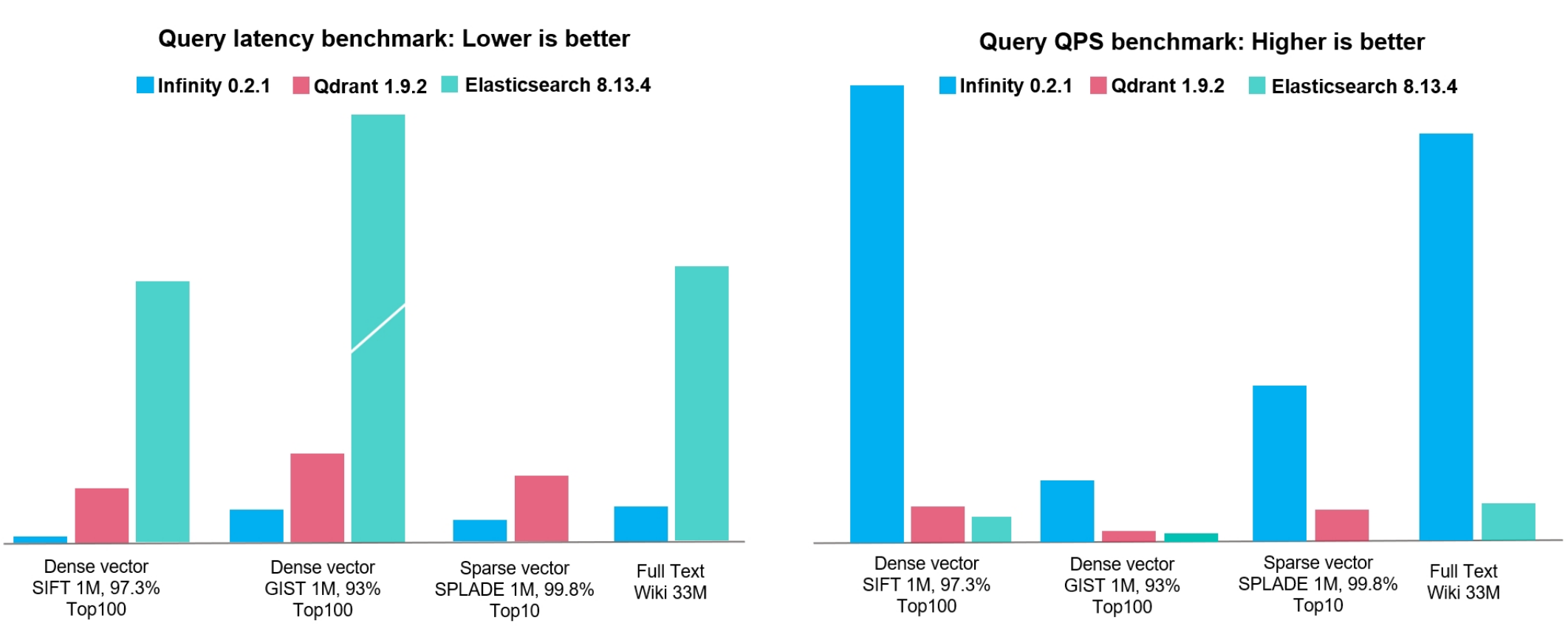
03 混合搜索

Indexing Database

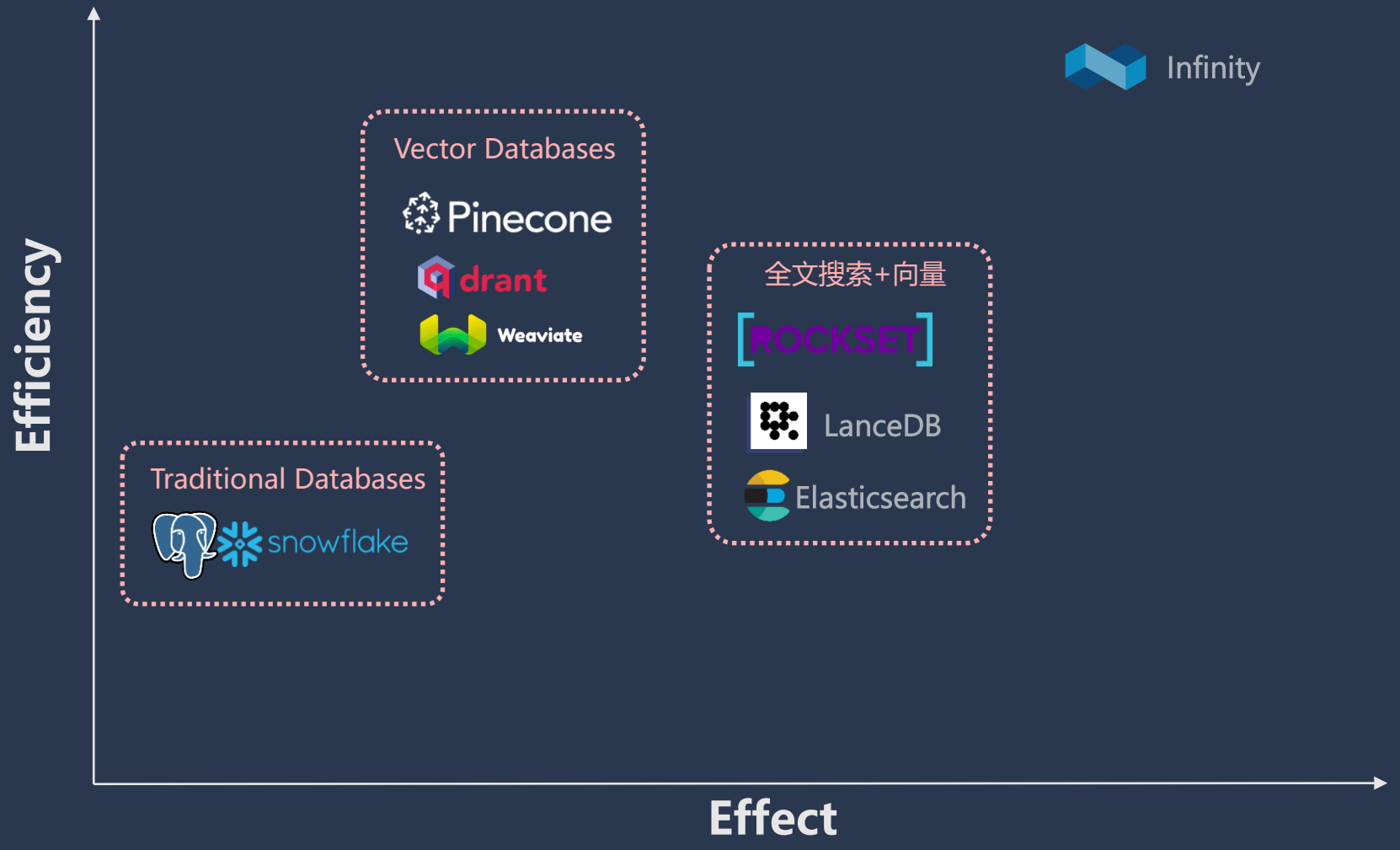




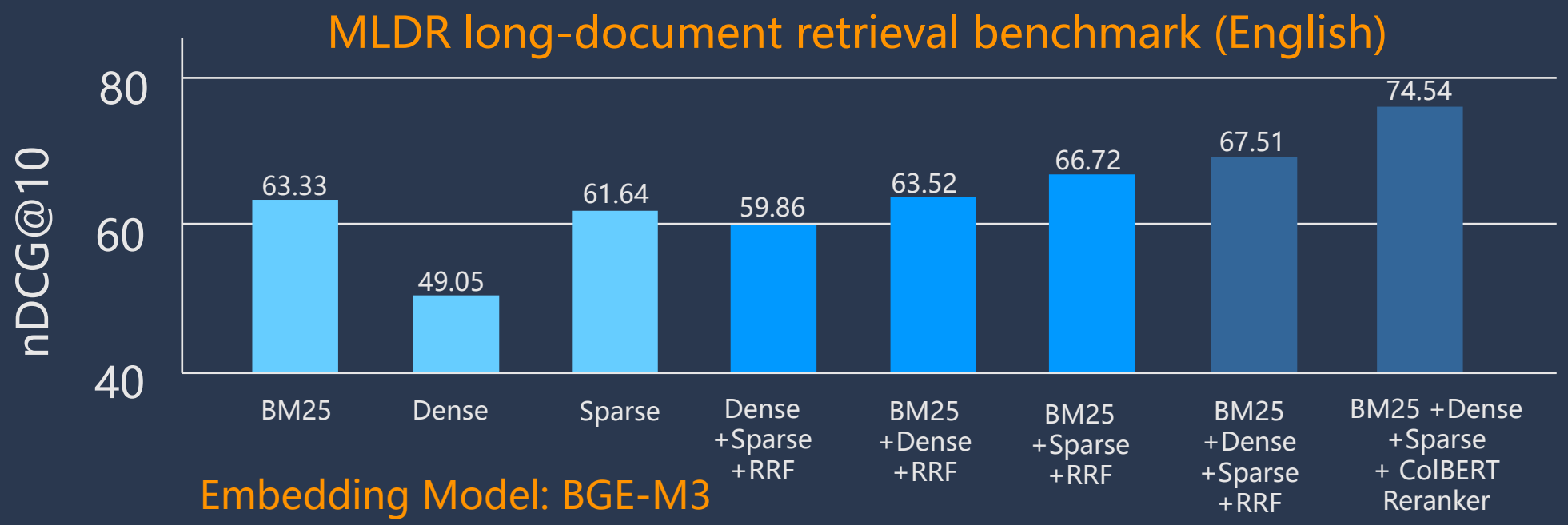
Benchmark


Query latency benchmark: Lower is better
Infinity 0.2.1
Qdrant 1.9.2
Elasticsearch 8.13.4
Dense vector SIFT 1M, 97.3% Top100
Dense vector GIST 1M, 93% Top100
Sparse vector SPLADE 1M, 99.8% Top10
Full Text Wiki 33M
Query QPS benchmark: Higher is better
Infinity 0.2.1
Qdrant 1.9.2
Elasticsearch 8.13.4
Dense vector SIFT 1M, 97.3% Top100
Dense vector GIST 1M, 93% Top100
Sparse vector SPLADE 1M, 99.8% Top10
Full Text Wiki 33M

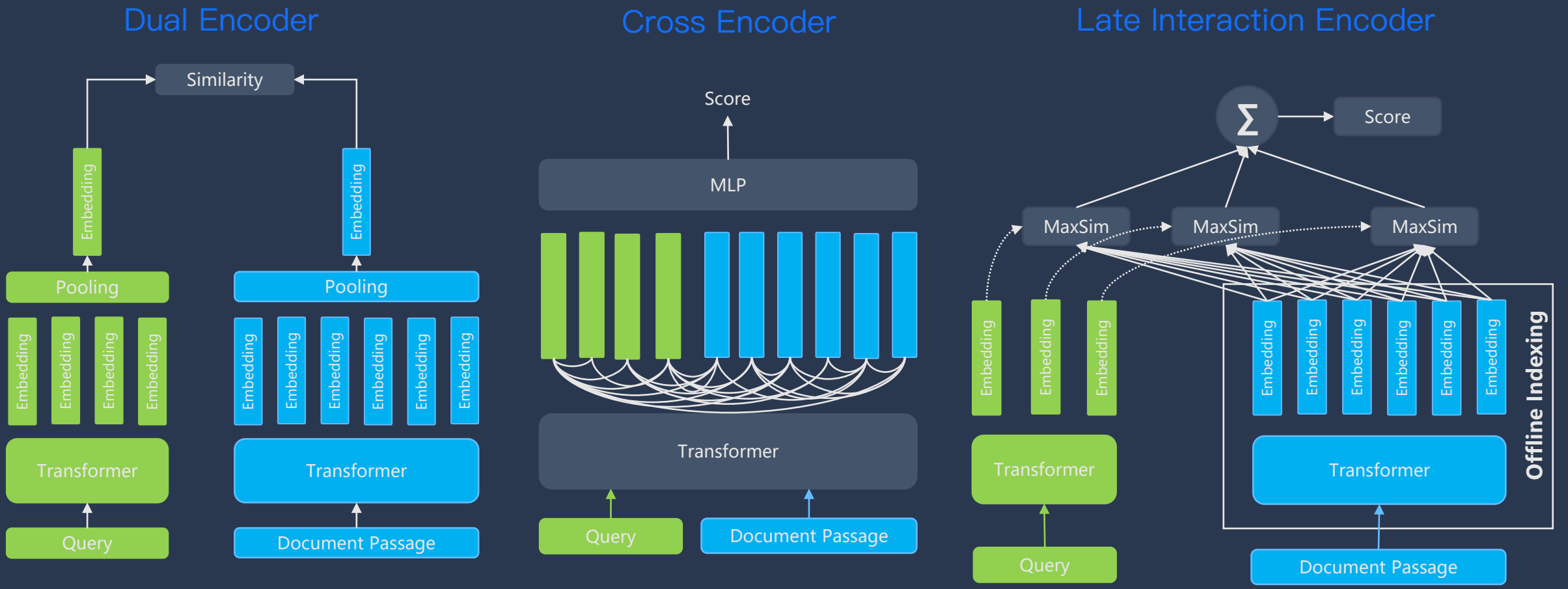
RAG数据库选型对比


Infinity
Vector Databases
Pinecone
drant
Weaviate
全文搜索+向量
ROCKSET
LanceDB
Elasticsearch
Traditional Databases
snowflake
Efficiency
Effect

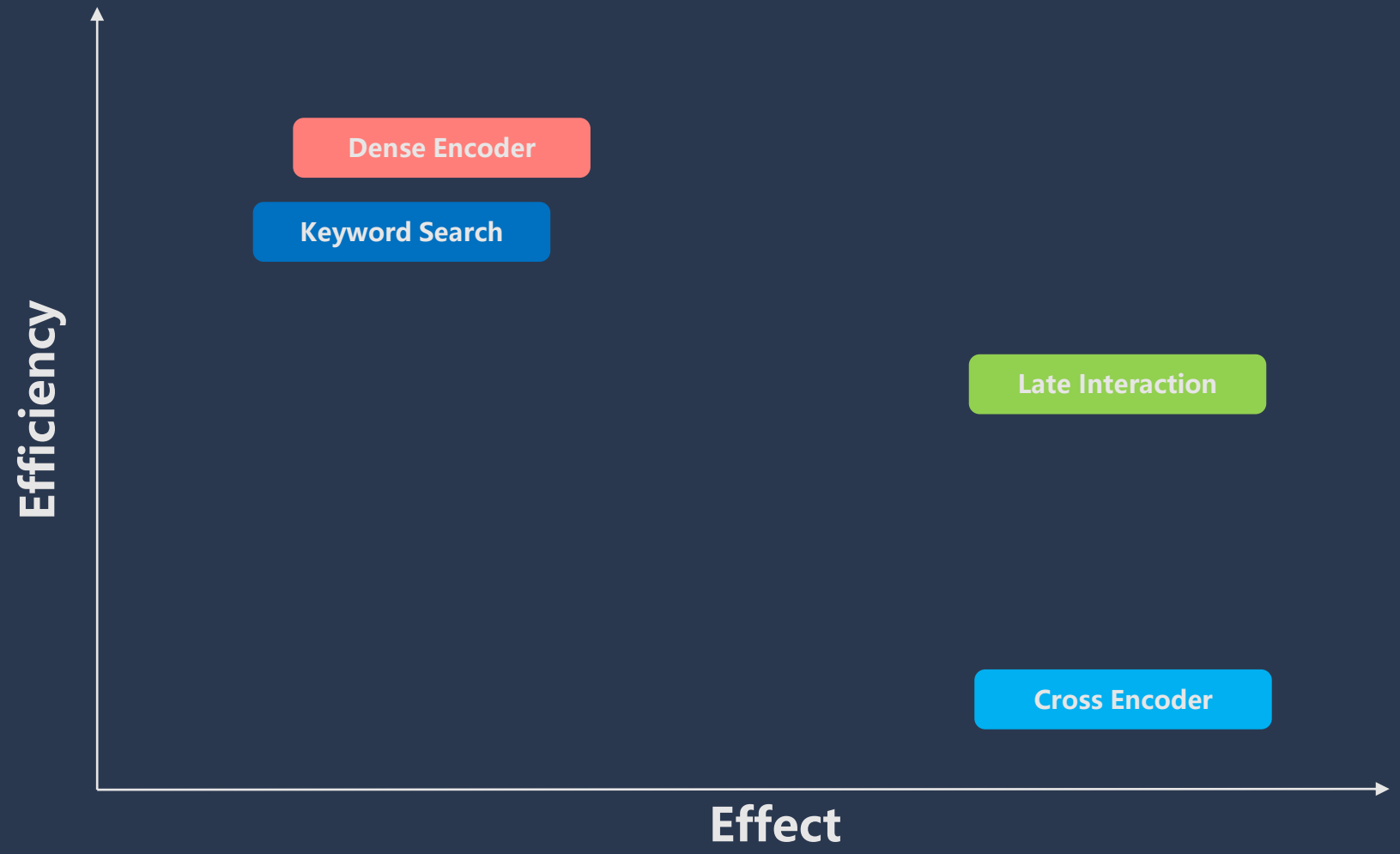
几路召回?


MLDR long-document retrieval benchmark (English)
nDCG@10
80
60
40
63.33
49.05
61.64
59.86
63.52
66.72
67.51
74.54
BM25
Dense
Sparse
Dense +Sparse +RRF
BM25 +Dense +RRF
BM25 +Sparse +RRF
BM25 +Dense +Sparse +RRF
BM25 +Dense +Sparse + ColBERT Reranker
Embedding Model: BGE-M3

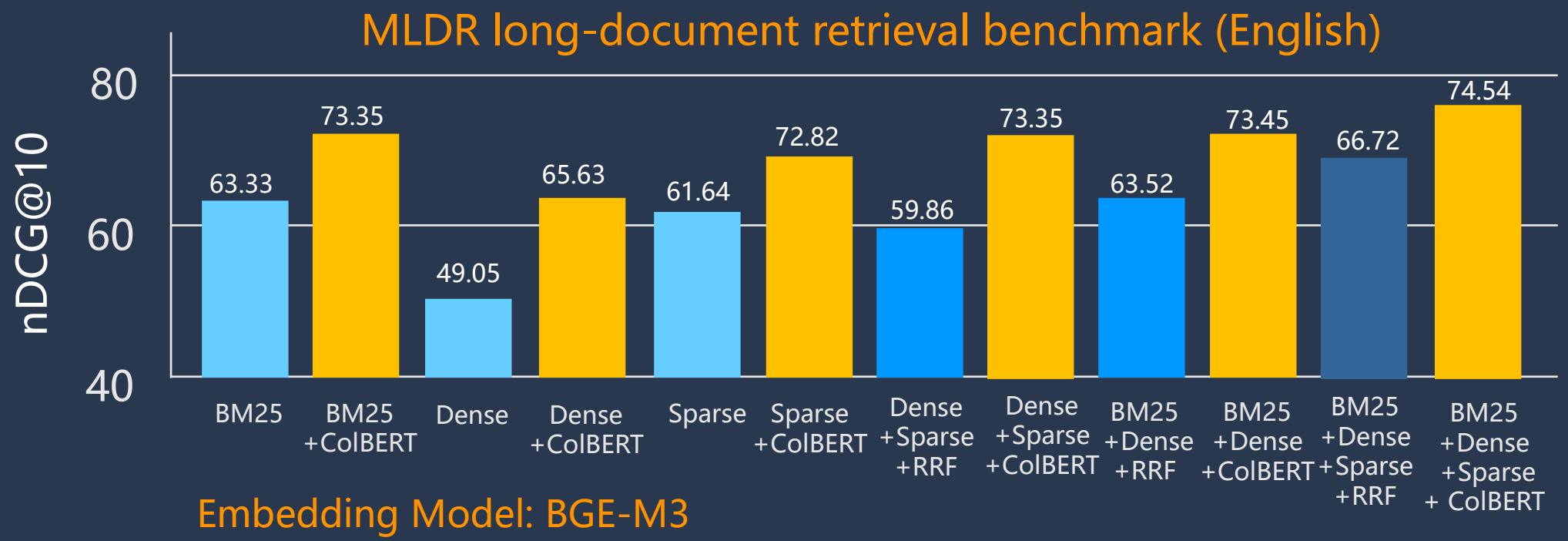
排序模型


Dual Encoder
Similarity
Embedding
Embedding
Pooling
Pooling
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Transformer
Transformer
Query
Document Passage
Cross Encoder
Score
MLP
Transformer
Query
Document Passage
Late Interaction Encoder
Σ
Score
MaxSim
MaxSim
MaxSim
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Transformer
Transformer
Query
Document Passage
Offline Indexing

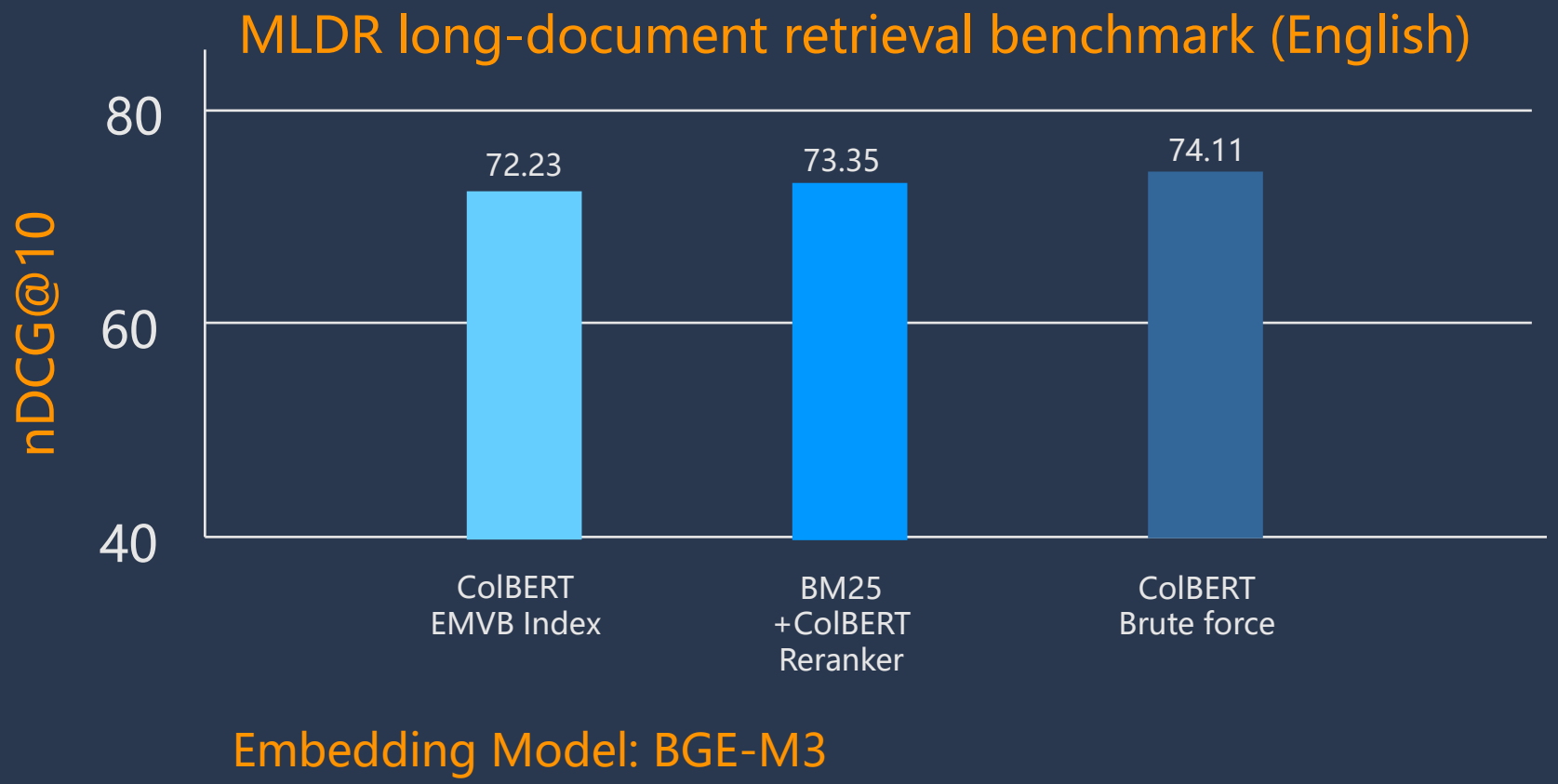
ColBERT的收益


Dense Encoder
Keyword Search
Late Interaction
Cross Encoder
Efficiency
Effect

ColBERT的收益

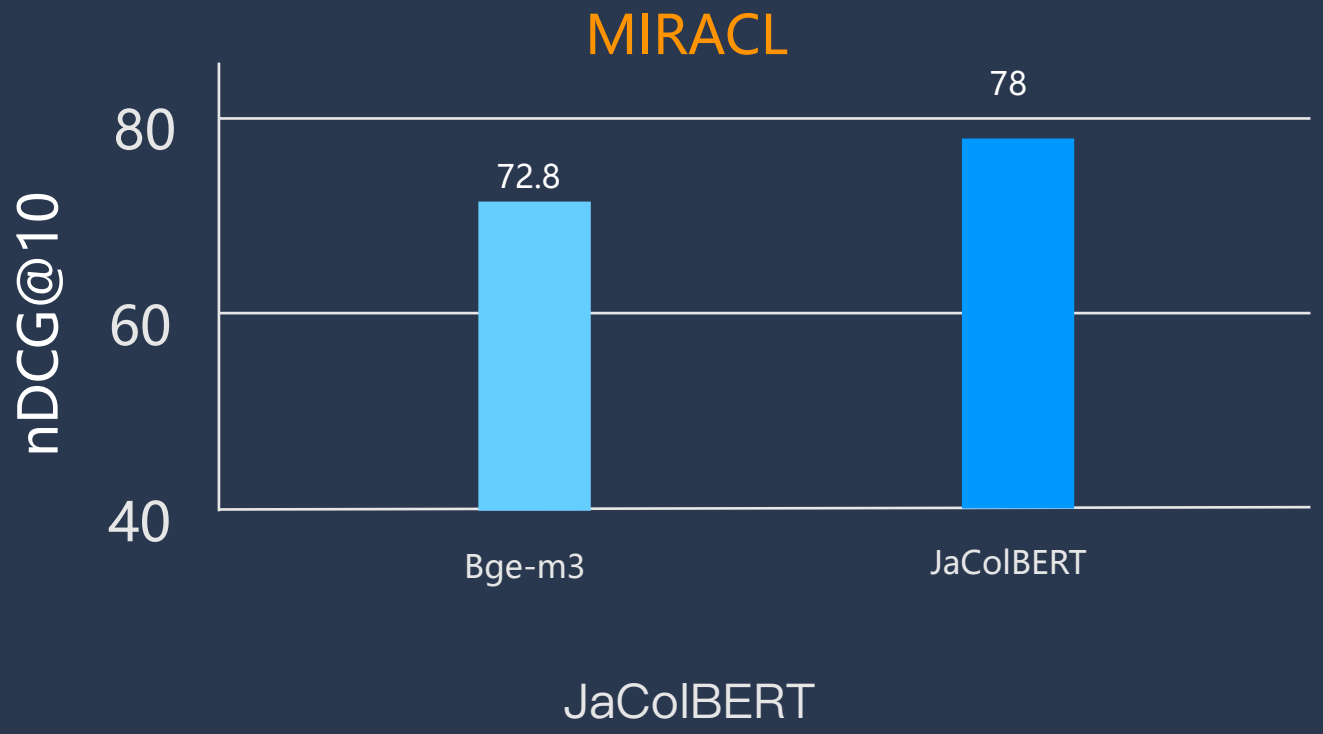

MLDR long-document retrieval benchmark (English)
nDCG@10
80
60
40
63.33
73.35
49.05
65.63
61.64
72.82
59.86
73.35
63.52
73.45
66.72
74.54
BM25
BM25 +ColBERT
Dense
Dense +ColBERT
Sparse
Sparse +ColBERT
Dense +Sparse +RRF
Dense +Sparse +ColBERT
BM25 +Dense +RRF
BM25 +Dense +ColBERT
BM25 +Dense +Sparse +RRF
BM25 +Dense +Sparse + ColBERT
Embedding Model: BGE-M3

ColBERT ranker 还是 reranker ?



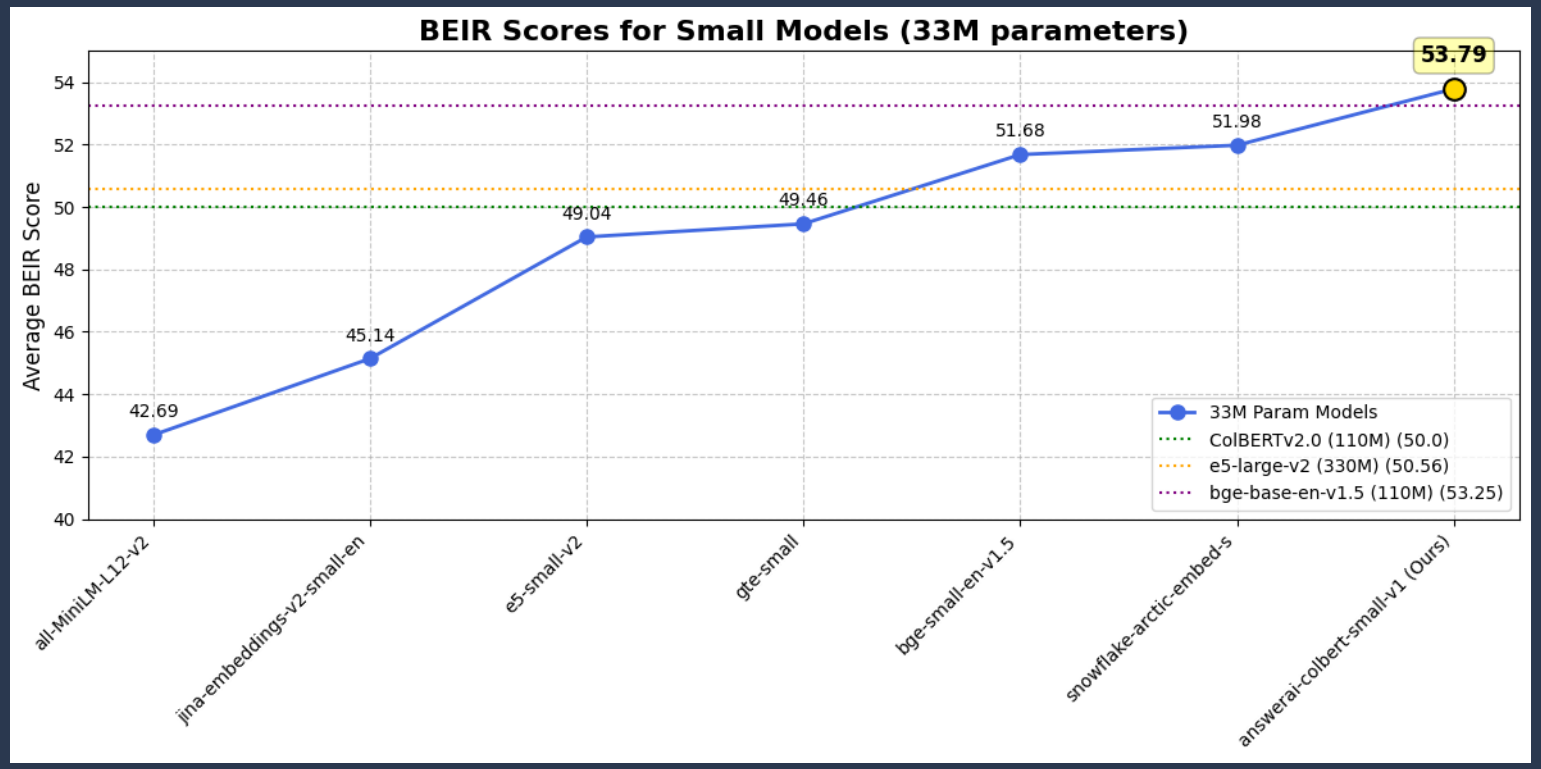
Embedding Model: BGE-M3

延迟交互是 RAG的未来



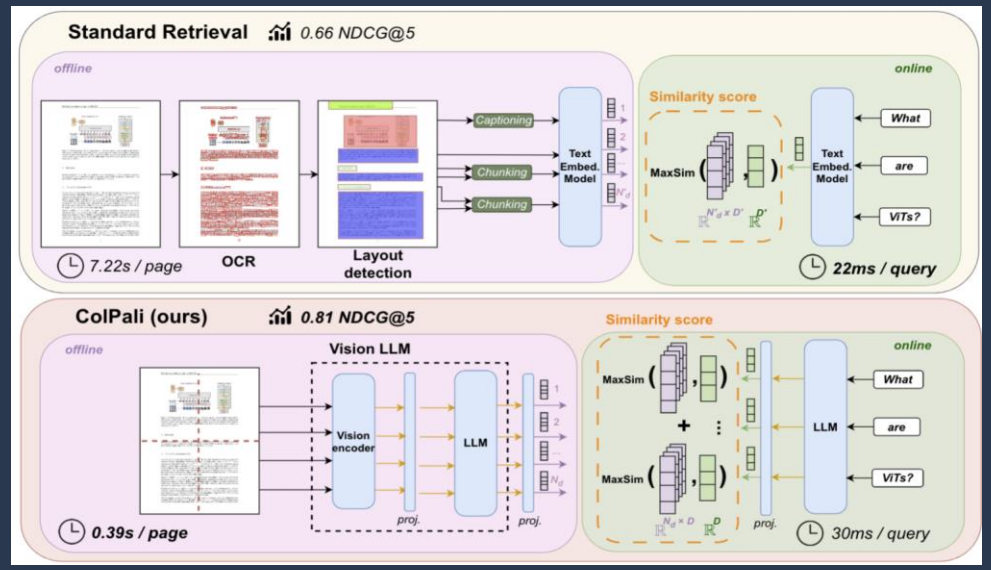
延迟交互是 RAG的未来

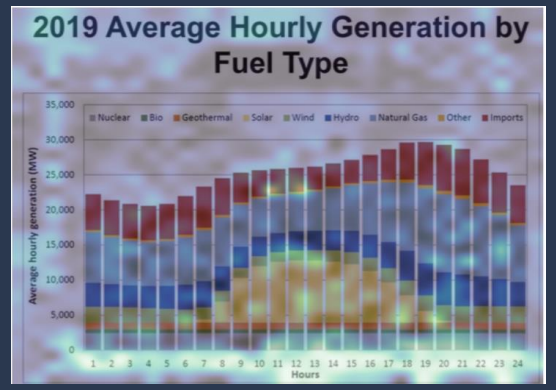
answerai-colbert-small-v1　基于JaColBERT　33M参数



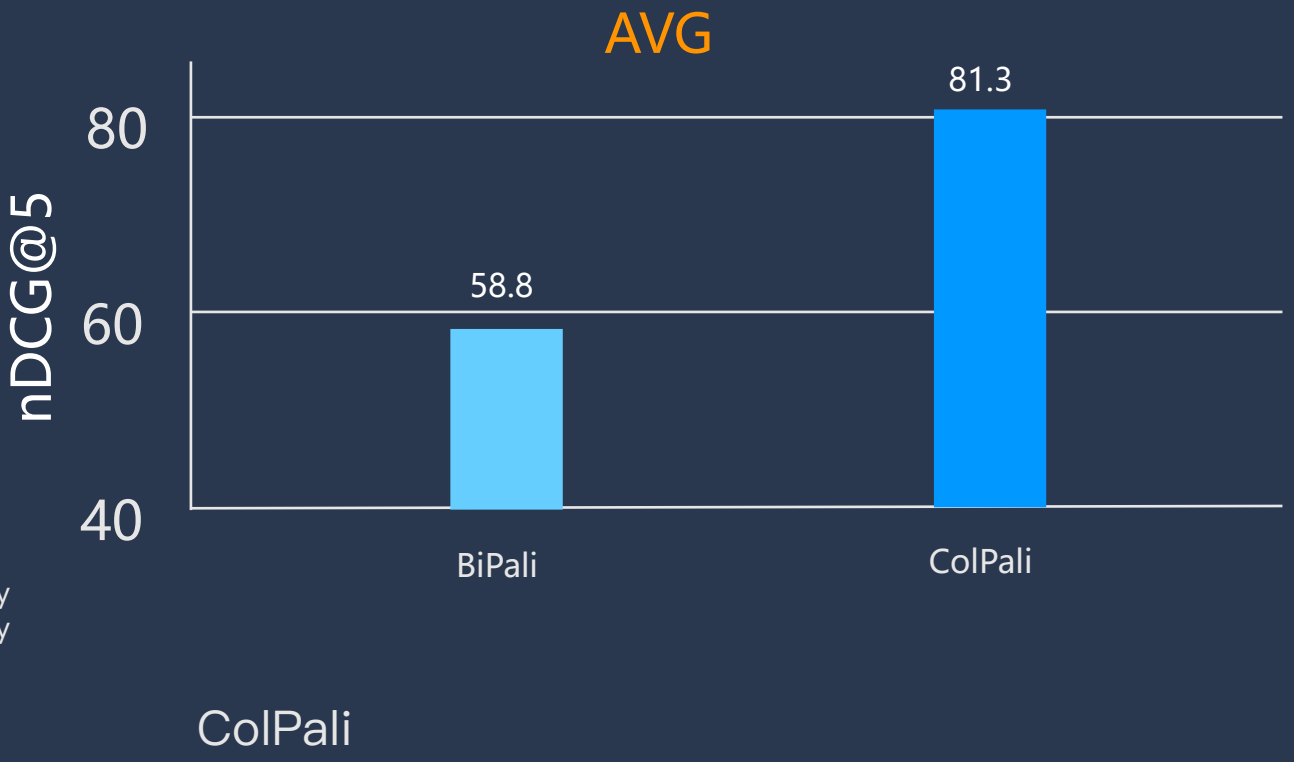
- 超过 BGE 110M
- 每个Token 96维
- Binary量化后每个Token 12 byte

延迟交互是 RAG的未来





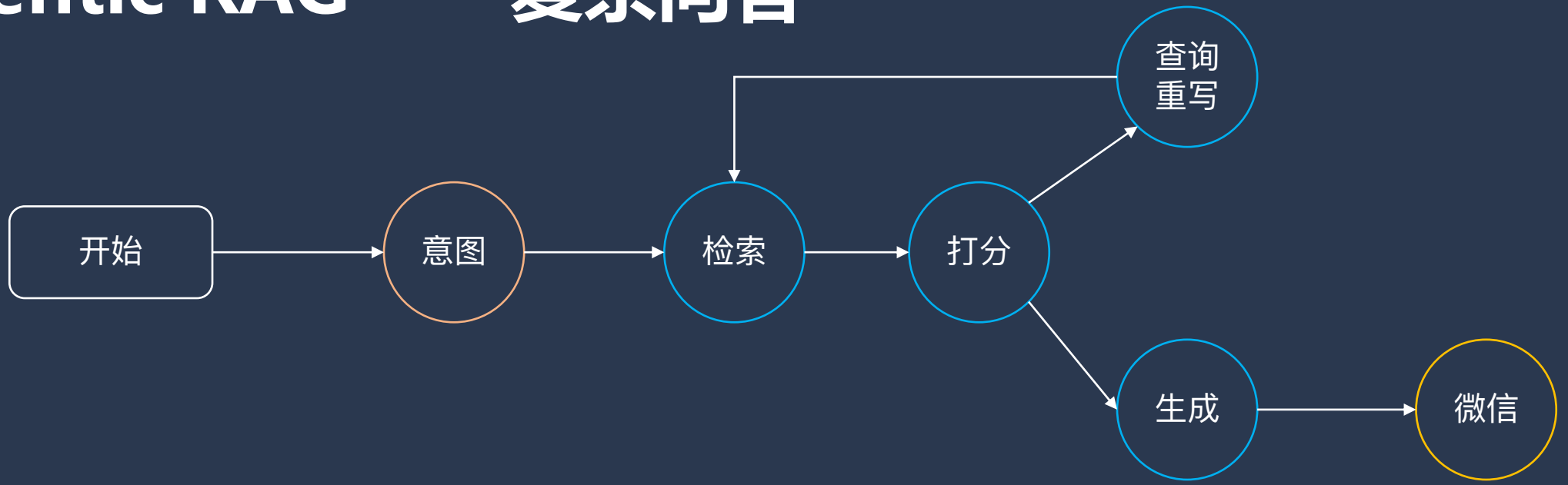
Query： Which hour of the day had the highest overall electricity generation in 2019 ?



ColPali

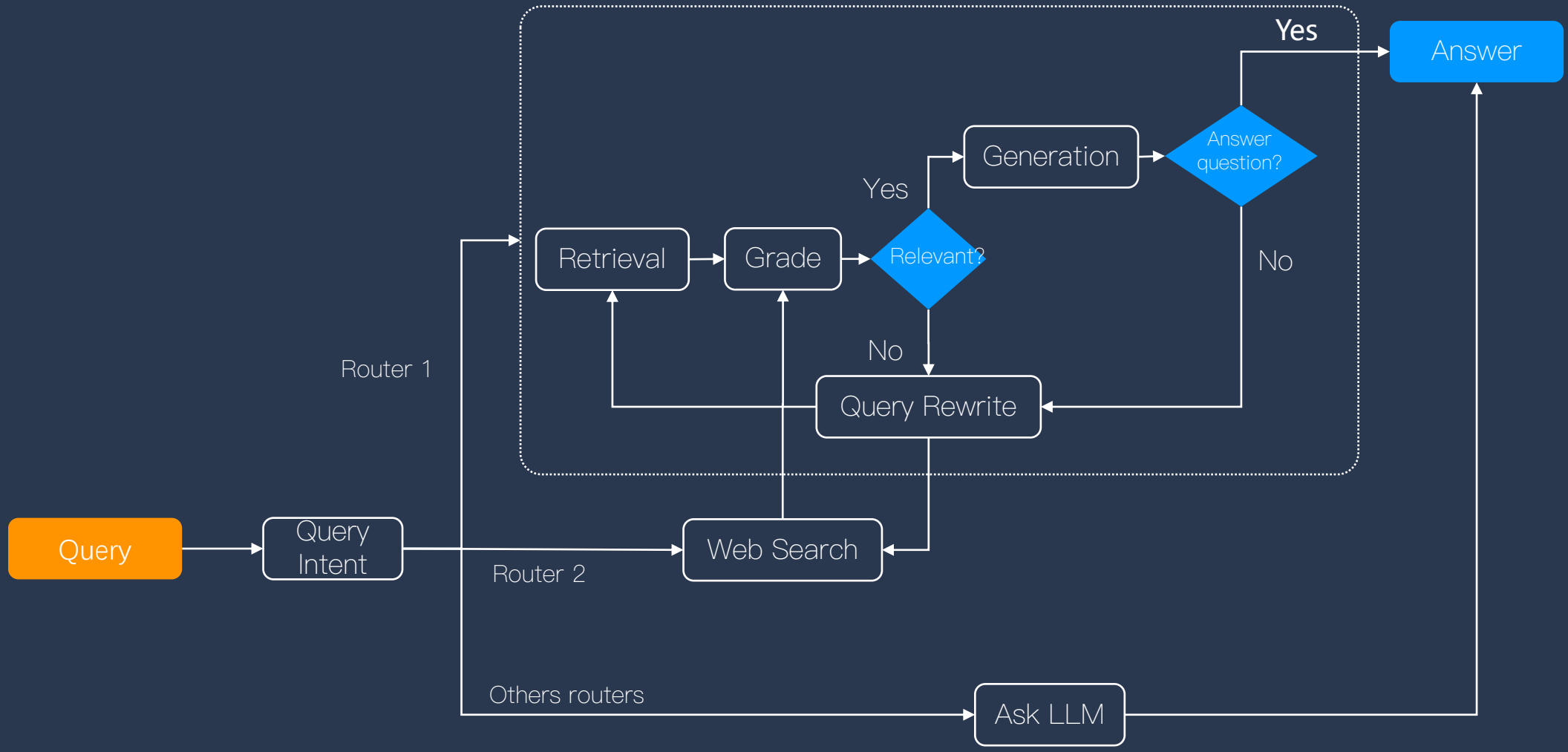
04 高级RAG

Agentic RAG——复杂问答

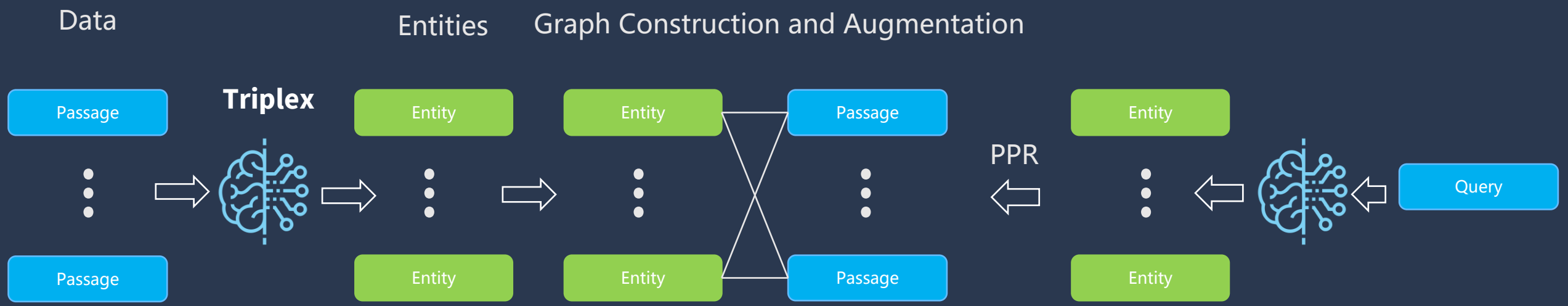


Agentic RAG	
✔ 反思——自我纠错和迭代 ✔ 工具——workflow ✔ 规划——workflow ✘ 多Agent协作	■ 查询意图识别 ■ 知识图谱 ■ 查询改写 ■ 文档解析算法 ■ 文档聚类和摘要RAPTOR ■ …

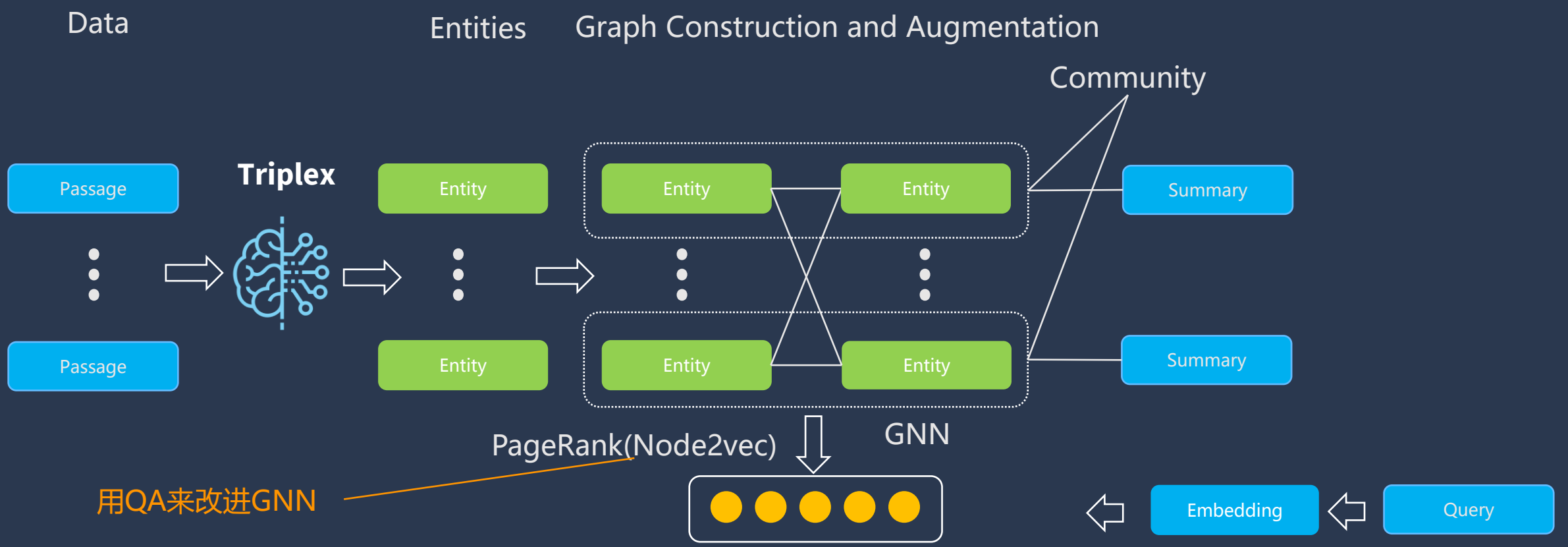
Agentic RAG——复杂问答



知识图谱


Data
Entities
Graph Construction and Augmentation
Passage
Passage
Triplex
Entity
Entity
Entity
Entity
Passage
Passage
PPR
Entity
Entity
Query

知识图谱


Data
Entities
Graph Construction and Augmentation
Community
Passage
Triplex
Entity
Entity
Entity
Summary
Passage
Entity
Entity
Entity
Summary
PageRank(Node2vec)
GNN
用QA来改进GNN
Embedding
Query

下一代RAG平台


电商
AI搜索
客服
人力
猎头
员工支持
合规
文档助手
合规校验
汽车
个性化服务
工作流支持
RAG 2.0 Platform
Agent 引擎
AI Native Database
文档理解
KG 构建
多模态Embedding
<3B
医疗
客户服务
个性化服务
制造
质量控制
运营提效
财务
客户服务
决策支持
IT
客户服务
审计自动化

InfoQ 极客传媒

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

4月 — QCon 北京 · 全球软件开发大会 · 会议时间：已结束

5月 — AiCon 北京 · 全球人工智能开发与应用大会 · 会议时间：已结束

6月 — ArchSummit 深圳 · 全球架构师峰会 · 会议时间：已结束

8月 — FCon 上海 · 全球金融科技大会 · 会议时间：8月16–17日

8月 — AiCon 上海 · 全球人工智能开发与应用大会 · 会议时间：8月18–19日

10月 — QCon 上海 · 全球软件开发大会 · 会议时间：10月18–19日

12月 — ArchSummit 北京 · 全球架构师峰会 · 会议时间：12月20–21日

THANKS

智能未来，探索 AI 无限可能

Intelligent Future, Exploring the Boundless Possibilities of AI

https://github.com/infiniflow/ragflow
https://github.com/infiniflow/infinity

AiCon
全球人工智能开发与应用大会